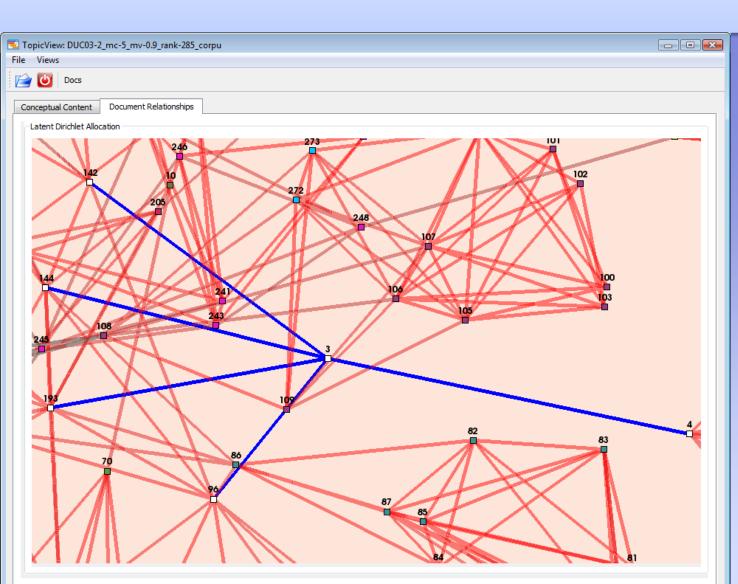
TopicView: Understanding Document Relationships Using Latent Dirichlet Allocation (LDA) Models

Patricia J. Crossno, Andrew T. Wilson, Daniel M. Dunlavy, and Timothy M. Shead (Sandia National Laboratories)

Unexpected Topic Connections

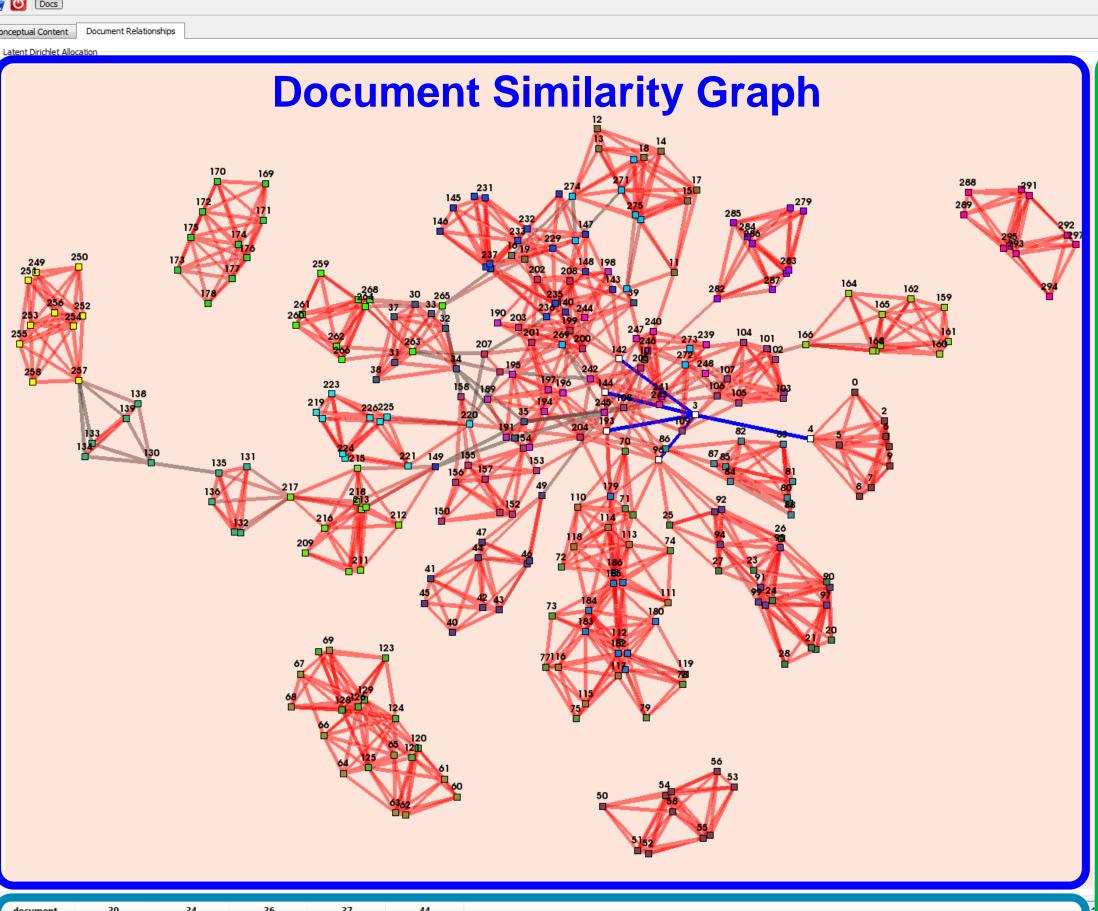


- LDA model of newswire articles from 2003
 Document Understanding Conference (DUC)
 contest has unexpected links between articles.
- For example, documents 3 and 4 are articles on Pinochet's arrest in Britain (topic 44), yet 3 is connected to articles on unrelated topics.
- Document 96 is an article describing Israel closing a Palestinian airport (topic 34).
- Documents 142 and 144 are articles about electing judges to a Yugoslav war crimes tribunal in the Hague (topic 36).
- Document 193 discusses cold weather deaths in Moscow (topic 37).

TopicView Model Exploration



- Term-Topic Table lists terms in order of importance within each topic. Term weight shown by text darkness.
- Pocument Similarity Graph displays document relationships. Edge weights color-coded red (high) to gray (low). Documents (nodes) colored by human generated topic groups provided in DUC contest data.
- Document-Topic Table provides document weights within each topic. Weight color-coded in text darkness.
- **Document Text** provides full text of selected documents with selected terms in Term-Topic Table shown in **red**.
- Document selection, highlighted by white nodes and blue edges in the graph, is linked between views.



Document-Topic Table

Text: APW19981017.0507 NEWS STORY 10/17/1998 12:48:00 w1371 &Cx1f; wstm-ui &Cx13; &Cx11; BC-Britain-Pinochet 6thLd 10-17 0105 BC-Britain-Pinochet, 6th Ld,0105 British police arrest Pinochet on Spanish warrant &UR; Eds: RECASTS top two grafs. &QL; &UR; By SUE LEEMAN &QC; Associated Press Writer &QC; LONDON (AP) _ Former Chilean dictator Gen. Augusto Pinochet has been arrested by British police on a Spanish extradition warrant, despite protests from Chile that he is entitled to diplomatic immunity. Spanish judges ar more... (POFILE (WS SLBC-Britain-Pinochet, 6th Ld; CT (REG:EURO;) (REG:BRIT;) (REG:SCAN;) (REG:MEST;) (REG:AFRI;) (REG:INDI; (REG:ENGL;) (REG:ASIA;) (LANG:ENGLISH;)))
AP-NY-10-17-98 1248EDT
Group: 0

Text: APW19981018.0410 NEWS STORY 10/18/1998 10:29:00 w2283 &Cx1f; wstm-ri &Cx13; &Cx11; BC-Ibero-AmericanSummit- 10-18 0438 BC-Ibero-AmericanSummit- 10-18 0438 BC-Ibero-American Summit-Castro-Pinochet Castro disagrees with arrest of Pinochet &UR; By VIVIAN SEQUERA &QC; &UR; Associated Press Writer &QC; OPORTO, Portugal (AP) _ President Fidel Castro said Sunday he disagreed with the arrest in London of former Chilean dictator Augusto Pinochet, calling it a case of international meddling." Castro compared the action with the establishment in Rome in August of an International Criminal Court, a move Cuba has expressed reservations about. ``It seems to me that what has happened there (in London) universal meddling," Castro told reporters covering the Ibero-American summit being held here Sunday. Castro told reporters covering the Ibero-American summit being held here Sunday. Castro told reporters covering the Ibero-American summit period by a Spanish in a city hotel. He said the case seemed to be ``unprecedented and unusu Pinochet, 82, was placed under arrest in London Friday by British police acting of a warrant issued by a Spanish judge. The judge is probing Pinochet's role in the death of Spaniards in Chile under his rule in the 1970s and 80s. Castro, Latin America's only remaining authoritarian leader, said he lacked details on the case against Pinochet, but said he thought it placed the government of Chile and President Eduardo Frei in an uncomfortable position while Frei is attending the summit. Castro said the International Criminal Court ought to be i more ... (PROF. (WS St.:BC-Ibero-American Summit-Castro-Pinochet, CT:; (REG:EURO;) (REG:SCAN;) (REG:ENGL;) (LANG:ENGLISH;))) AP-NY-10-18-98 1029EDT Group: 0

Text: APW 19981230.0991 NEWS STORY 12/30/1998 17:16:00 w 1336 &Cx1f; wstm-ri &Cx13; &Cx11; BC-Israel-Palestinians 1stLd 12-30 0063 BC-Israel-Palestinians, 1stLd Israel delays two flights from Palestinian airport & Eds: RESTORES dropped word `marking' from lede &QL; JERUSALEM (AP) _ Isr security officials delayed two planes from taking off from the Palestinian airport Wed more ... (PROFILE (WS SL:BC-Israel-Palestinians, 1stLd; CT:i; (REG:EURC (REG:BRIT;) (REG:SCAN;) (REG:ENGL;) (REG:MEST;) (LANG:ENGLISH;))) AP-NY-12-30-98 1716EST Group: 9

ID: 142
Text: APW19981016.0667 NEWS STORY 10/16/1998 12:01:00 w2075 &Cx1f; wstm- u i &Cx13; &Cx11; BC-UN-YugoslavTribunal 1stt.d 10-16 0237
BC-UN-Yugoslav Tribunal, 1st Ld,0236 General Assembly votes to select three new judges for Hague tribunal &Qt; &UR; Eds: LEADS to UPDATE with no one getting majority in first round; Ecuador and Poland withdrawing nominations, narrowing field to seven. &Qt; &UR; By NICOLE WINFIELD &QC; &UR; Associated Press Writer &QC; UNITED NATIONS (AP) _The General Assembly voted Friday to select three judges to fill a new trial chamber of the Yugoslav war crimes tribunal, whose caseload could grow with atrocities committed during the Kosovo conflict. Representatives of the 185-member assembly cast secret ballots, choosing among nine candidates nominated by their home countries. The three will rule on cases concerning crimes committed in the former Yugoslavia. Their terms will end Nov. 16, 2001. None of the original nine candidates emerged from the more.... (PROFILE (WS SL:BC-UN-Yugoslav Tribunal, 1st Ld; CT:j; (REG:EURO;) (REG:BRIT;) (REG:SCAN;) (REG:MEST;) (REG:AFRI;) (REG:INDI;) (REG:ENGL;) (REG:ASIA;) (LANG:ENGLISH;)))

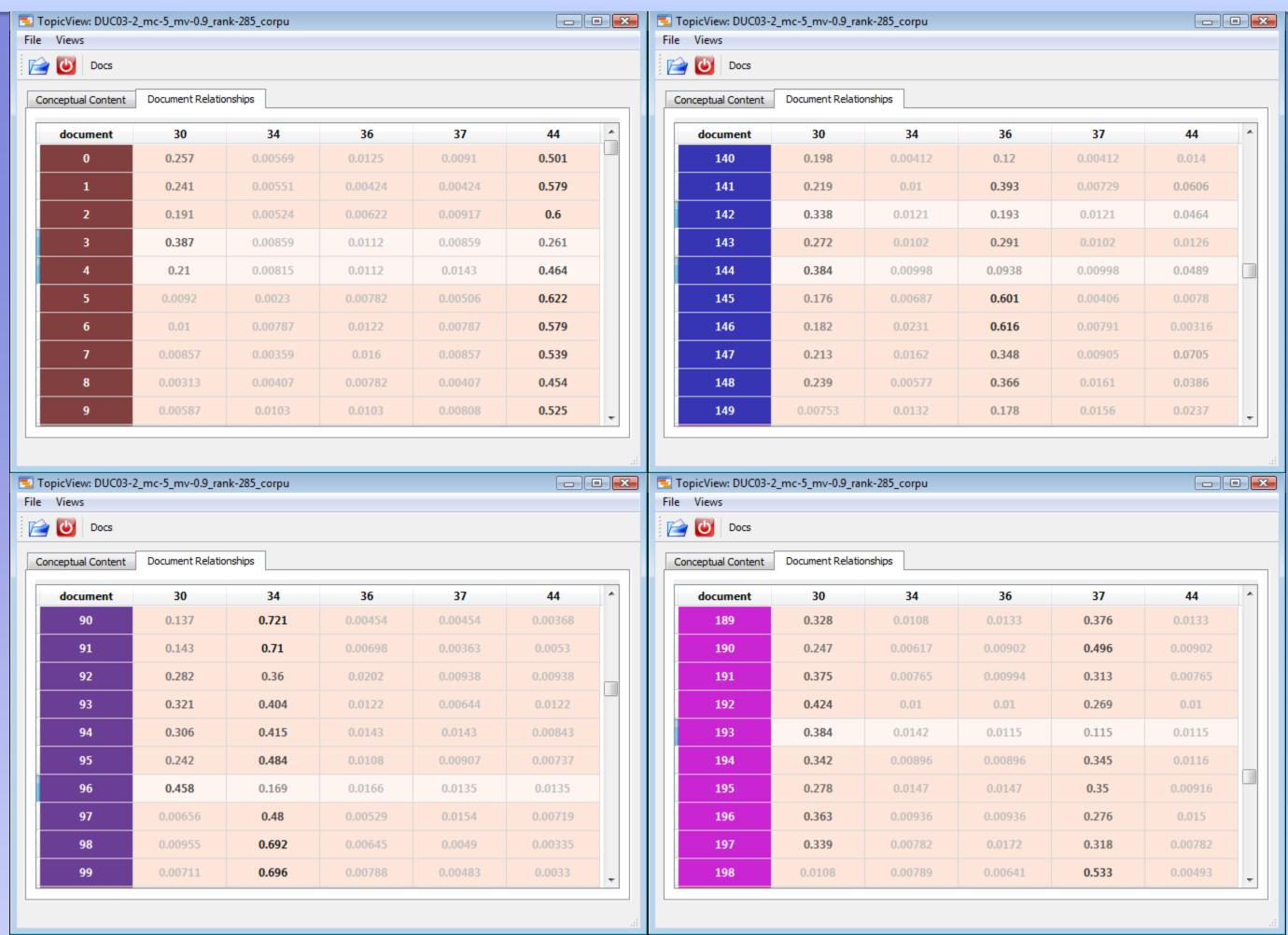
AP-NY-10-16-98 1201EDT Group: 14

Text: APW19981029.0281 NEWS STORY 10/29/1998 05:19:00 w2425 &Cx1f; wstm-ui &Cx13; &Cx11; BC-Bosnia-WarCrimes 10-29 0092 BC-Bosnia-War Crimes Bosnian Serb pleads guilty to 14 murders at U.N. tribunal &UR; URGENT &QL; THE HAGUE, Netherlands (AP) Handing the Yugoslav war crimes tribunal only its third conviction, a Bosnian Serb Thursday pleaded guilty 14 murders. Goran Jelisic, 30, faces a maximum life prison term for more... (PROFILE (WS SL:BC-Bosnia-War Crimes; CT:i; (REG:EURO;) (REG:BRIT;) (REG:SCAN;) (REG:BRGL;) (LANG:ENGLISH;))) AP-NY-10-29-98 0519EST Group: 14

ID: 193
Text: APW19981130.0222 NEWS STORY 11/30/1998 05:14:00 w2620 &Cx1f; wstm-ri &Cx13; &Cx11; BC-Russia-Cold-Deaths 11-30 0139
BC-Russia-Cold-Deaths Cold kills 39 in Moscow this month MOSCOW (AP) _
Unusually cold weather in Moscow killed at least 39 people in November, medical officials said on Monday. Winter has come early to the Russian capital this year, with temperatures below freezing through most of the month. Since the beginning of October, 45 people have died of hypothermia, the city ambulance service said, accomore... (PROFILE (WS SL:BC-Russia-Cold-Deaths; CT:i; (REG:EURO;) (REG:BRIT;) (REG:SCAN;) (REG:MEST;) (REG:AFRI;) (REG:INDI;) (REG:ENGL;) (REG:ASIA;) (LANG:ENGLISH;)))
AP-NY-11-30-98 0514EST
Group: 19

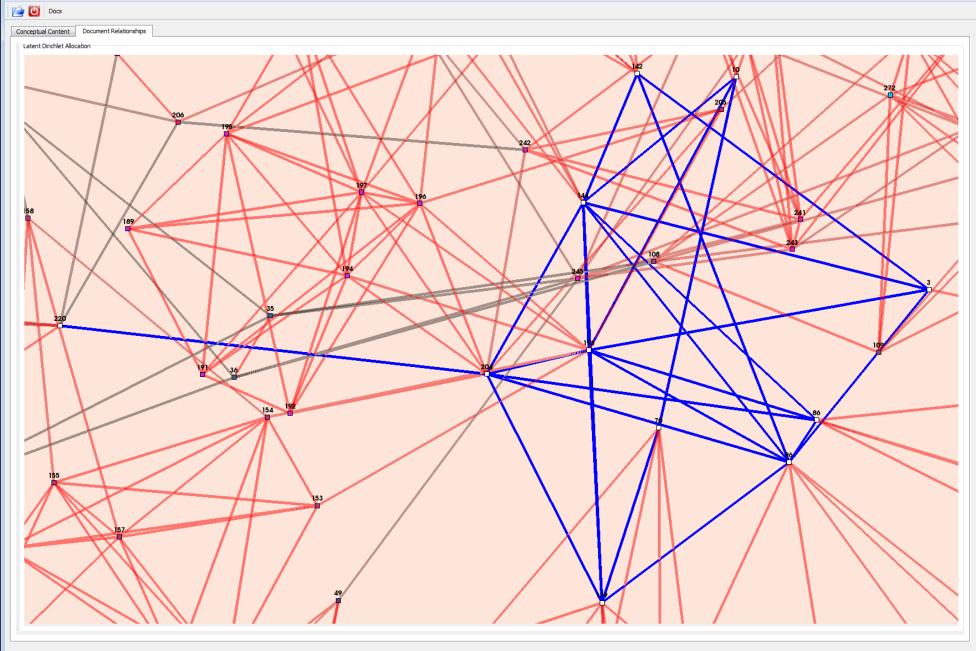
Document Text

Document-Topic Weight Pattern



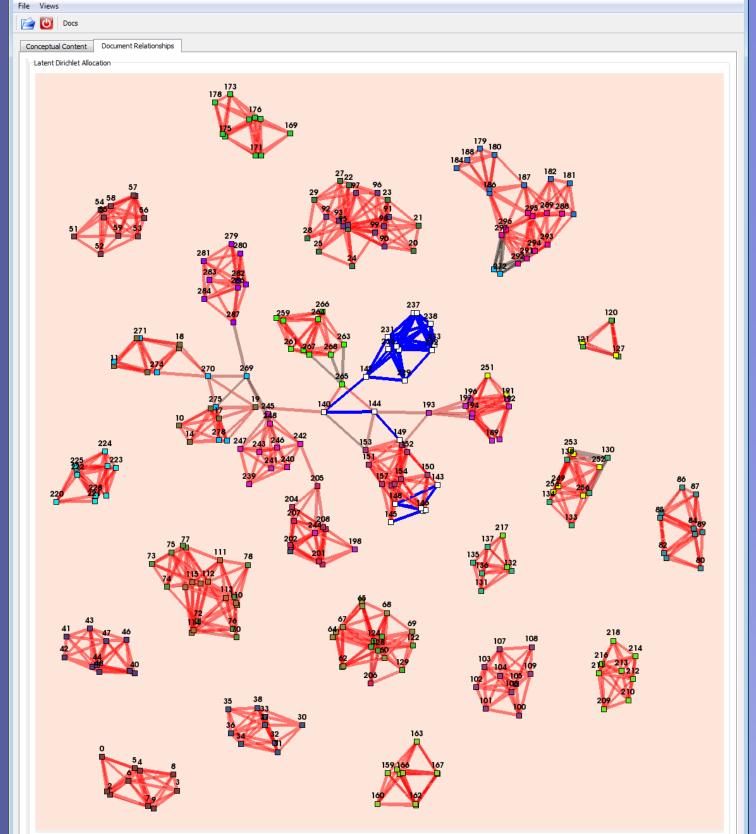
- Document -topic weights for documents 3, 4, 96, 142, 144, and 193 are shown in rows with pale backgrounds for topics (columns) 30, 34, 36, 37 and 44.
- **Topic 30** is included because many of these documents are strongly weighted against it. Topic 30's most significant terms consist of XML header tags used in articles from the **Associated Press**. (Topic 31, not shown, similarly groups articles from the New York Times).
- An interesting pattern emerges (for all articles except 4, which is connected to 3 by concept) the articles are more strongly weighted in their source topic than in the topic representing their conceptual content.
- Hypothesis: Documents matching this pattern are the source of many of the edges between disjoint topics.

Bridging Documents



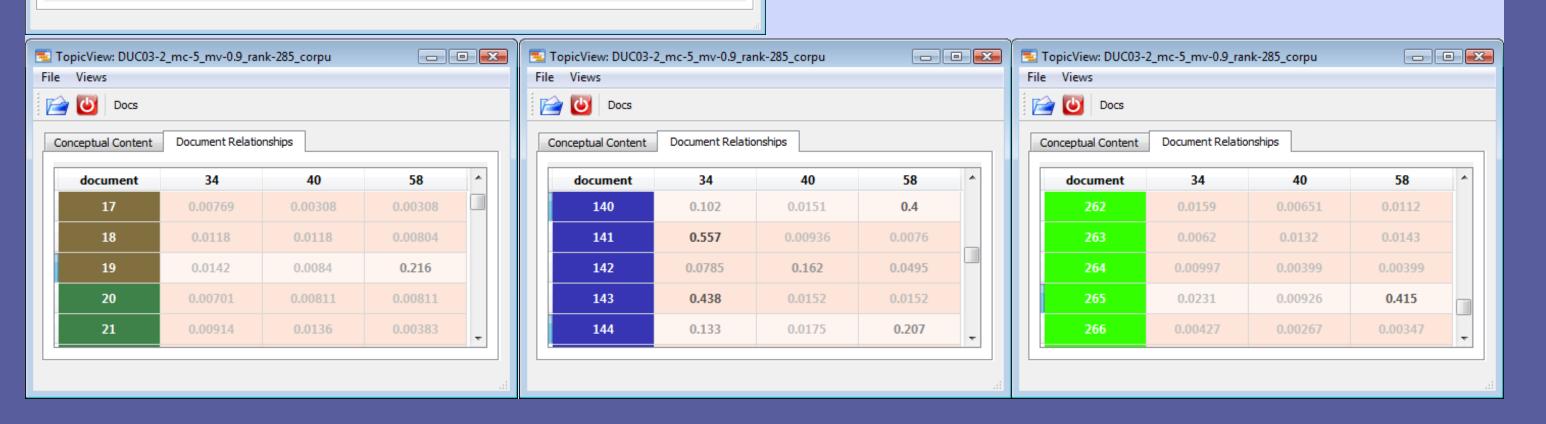
- We test our hypothesis by listing all documents whose strongest weights are AP or NYT topics.
- For each document in the list, we check all linked documents in the graph (like we did for 3).
- If linked documents share conceptual content (like 4), then the document fails the hypothesis.
- Of 297 documents, 33 fit the hypothesis. Of these, only 11 survived the test, having links not supported by conceptual content (including 3, 96, 142, 144, and 193).
- These 11 are in the center of the graph and tend to link with one other, impacting layouts of their associated clusters. We call them **bridging documents**. They are **all short AP articles**, where the terms in the headers outweigh the text in the news content (NYT articles tend to be longer).
- Revised hypothesis: Only documents whose conceptual content is outweighed by the source content will display this bridging pattern that links disjoint topics.

Merged Topics & Bridging Topics



- We test our revised hypothesis by rerunning LDA on headlines and story bodies only.
- Many edges linking disjoint topics disappear.
- Document clusters separate and layout
- AP & NYT topics disappear, freeing 2 topics.

 DUC data consists of 30 labeled topic groups.
- Previous topic 36 splits into new topics 34 & 40, separating Yugoslav war crimes tribunal and Iranian election stories (highlighted by white nodes and blue edges in the upper graph).
- Now articles on Chechen kidnappings have been merged with Yugoslav war crimes tribunal articles in topic 34.
- Document 142 still links topics 34 and 40. As with the AP source, 142 is more strongly linked to topic 40 (Iranian elections) than to its conceptual topic, 34 (tribunal-kidnappings).
- In topic 40, assembly and candidates are the 2nd and 3rd strongest terms. In document 142, assembly and candidates appear 3 and 4 times each. These terms dominate the other terms and strongly link 142 to topic 40.
- Document 142 remains a bridging document, though the conceptual link is valid here.
- Narrowly-focused topics appear, i.e. topic 58 linking trial-related stories (shown in the lower graph linked by blue edges, documents 19, 140, 144, and 256). This bridging topic (similar to the AP topic) connects documents more strongly than their conceptual groups.
- Choice of topic count impacts clustering, given merging and splitting of topics. Tried counts from 28 to 78. Correct count unclear.
- More topics will not necessarily separate combined topics or reduce edges. Topics can merge and bridging topics appear.
- With more topics, document weights within some topics may become so low that they appear to be **noise**.



Contact Information:

Patricia J. Crossno: pjcross@sandia.gov
Andrew T. Wilson: atwilso@sandia.gov
Daniel M. Dunlavy: dmdunla@sandia.gov
Timothy M. Shead: tshead@sandia.gov

Acknowledgements:

This work was funded by the Laboratory Directed Research & Development (LDRD) program at Sandia National Laboratories.

